DOCUMENT RESUME

ED 036 856 24 CG 005 117

AUTHOR Jones, Margaret Hubbard
TITLE Reliability of Coding of the System for the Analysis of Classroom Communication (SACC). Center for the Study of Evaluation Working Paper Number Two.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY Office of Education (DHEW), Washington, D.C. Bureau of Research.
BUREAU NO BR-6-1646
PUB DATE Nov 69
NOTE 27p.

EDRS PRICE EDRS Price MF-$0.25 HC-$1.45
DESCRIPTORS *Classroom Communication, *Classroom Observation Techniques, Communication (Thought Transfer), Evaluation, Measurement Instruments, Measurement Techniques, Observation, *Reliability, *Students, *Teachers, Verbal Communication

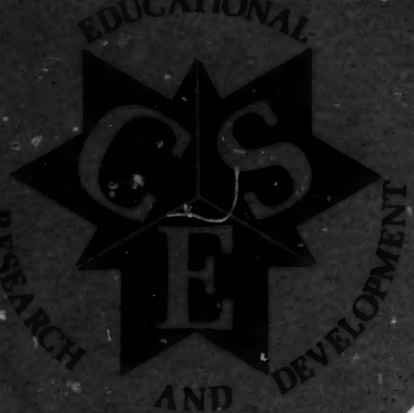
ABSTRACT

This study assesses the reliability of the System for the Analysis of Classroom Communication (SACC) devised to permit the gathering of data descriptive of classroom communication between teacher and pupils for evaluative purposes. The reliability used was called inter-observer agreement. The measure of inter-observer agreement used was the Scott coefficient which takes into account the number of categories in the system and the frequency with which each is used. The sample consisted of six schools, 20 teachers, eight subject matters, and eight grade groups. Students were of average socioeconomic status, most were "Anglos" but some were Mexican-Americans. There were 33 sessions varying in length from seven to 34 minutes; a session being defined as a coherent curricular unit. The two observers, who coded at the same sessions, were advanced graduate students in education. Results indicated that the level of inter-observer agreement was significantly high enough to permit use of the instrument for evaluative purposes. A modification of procedure should be used when the goal is evaluation of a school or a grade level. (EK)

ED036856

OE-BR 6-1646
PA 24
W.P. # 2
CG

CENTER FOR THE
STUDY OF
EVALUATION
UCLA
Graduate School
of Education
Los Angeles, California



RELIABILITY OF CODING OF THE SYSTEM FOR THE ANALYSIS OF CLASSROOM COMMUNICATION (SACC)

Margaret Hubbard Jones

CSE Working Paper #2
November 1969

Center for the Study of Evaluation
UCLA Graduate School of Education

CG005117

CENTER FOR THE STUDY OF EVALUATION

Marvin C. Alkin, Director
Publications Committee:
James W. Trent, Chairman
Theodore R. Husek
Sherman J. Pearl
Audrey Schwartz

UCLA Graduate School of Education

The CENTER FOR THE STUDY OF EVALUATION is one of nine centers for educational research and development sponsored by the United States Department of Health, Education and Welfare, Office of Education. The research and development reported herein was performed pursuant to a contract with the U.S.O.E. under the provisions of the Cooperative Research Program.

Established at UCLA in June, 1966, CSE is devoted exclusively to finding new theories and methods of analyzing educational systems and programs and gauging their effects.

The Center serves its unique functions with an inter-disciplinary staff whose specialties combine for a broad, versatile approach to the complex problems of evaluation. Study projects are conducted in three major program areas: Evaluation of Instructional Programs, Evaluation of Educational Systems, and Evaluation Theory and Methodology.

This publication is one of many produced by the Center toward its goals. Information on CSE and its publications may be obtained by writing:

Office of Dissemination
Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California 90024

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

ED036856

RELIABILITY OF CODING OF THE SYSTEM FOR THE ANALYSIS OF CLASSROOM COMMUNICATION (SACC)

Margaret Hubbard Jones

CSE Working Paper #2
November 1969

Center for the Study of Evaluation
UCLA Graduate School of Education

I. INTRODUCTION

The System for Analysis of Classroom Communication (SACC) was devised to permit the gathering of data descriptive of classroom communication between teacher and pupils for general evaluative purposes. It is clear from many educational studies of different sorts (e.g., Bond & Dykstra, 1967) that "something" in the classroom affects pupil achievement. It is suggested by a number of studies of classroom interaction that different kinds of communication processes may result in different pupil outcomes. Hence it appears to be important to evaluate the classroom communication processes for any broad evaluation program. Many systems have been proposed for the purpose of describing classroom interaction, twenty-six of which are brought together in *Mirrors for Behavior*, (Simon & Boyer, 1967). These and other systems were reviewed for possible application to the evaluation problem. Because of unreliability, incompleteness, complexity, or cost, however, none of them proved to be suitable. Upon the bases of both theory and empirical results, a new instrument--the SACC--was devised, which was intended to be somewhat more analytical than the

simplest (and most used system--Flander's), yet less costly than the most complex. It has undergone five revisions, based upon experience in using the system in live classrooms. This development and the justification for it will be described in detail in a subsequent report. Attached is a copy of SACC, Form V (Appendix A).

In the summer of 1969 an attempt was made to assess the reliability of the instrument. Reliability is not a simple concept in this sphere. It is obviously desirable to have an observer capable of replicating his own coding, but this requires either typescript, audio tapes, or video tapes, and even the best of these provide less information than is available in the live classroom. Furthermore, the situations in which the permanent records are attained are ordinarily more constrained than a normal classroom. The technique mentioned was used in training the coders, but estimates of reliability in live situations were desired. Here two alternatives present themselves: a given coder can code two sessions with a single teacher and a single subject-matter, or two coders can code the same sessions. In the first case the question arises whether it is the teacher-consistency or the observer-reliability that is

largely responsible for discrepancies in the results. Although the literature seems to indicate strongly that teachers do not (in fact, cannot) change their style significantly without intensive training, it is quite possible that two very different sorts of lessons might occur--one where the pupils were largely learning certain tools, and a second where they were being encouraged to use the tools to arrive at new conclusions. Whereas such problems could be resolved with the teachers, the result would be increased stress upon the teacher, or greatly increased observation time, neither of which is desirable for an evaluation program. With proper sampling procedures, in a large-scale evaluation study, this would, in fact, create no difficulties, but it would for a briefer study of reliability. The second option, using two coders at the same session, was chosen as being most economical. This type of reliability is best called inter-observer agreement.

II. PROCEDURES

A brief description of the coding process will be useful here (see SACC, Form V, attached). SACC is a category system; all communicative behavior can be coded

into mutually exclusive categories. There are 12 major dimensions, 5 referring to teacher behavior, 5 referring to pupil behavior, and 2 which refer to either or both. Within the major dimensions are varying numbers of sub-categories, the number depending upon the kinds of distinctions that coders have been able to make, since the finer break-downs of earlier forms have proved unreliable. The total number of categories is 31, with four additional symbols used for special situations. Three of these last four are essential in studying inter-observer agreement, in order to keep the two records in step with one another; the fourth is a code-modifier to permit an estimate of the length of pupils' contributions. The system is committed to memory by the coder, and coding is practiced on several kinds of materials until the coder can reproduce to a reasonable degree the "master" coding and until his speed has increased to the point where he can code at classroom pace.

Coding is done every 5 seconds, paced by a timer (see Apparatus Report) which actuates both a buzzer and a light, as well as displaying the number of the cell to be coded. If there is a change in major dimension within the 5-second interval, both codes are entered

in the same cell (occasionally three codes are required). The difficulty experienced in practice when two individuals judge the time at which events occur is a very old problem, going back to 1796 when the Astronomer Royal of Britain dismissed his assistant for "errors" in observation of the transit of stars, which led to a series of researches on "prior entry." (Boring, 1950) The implication of this research is that the time at which an event is observed to occur is a function of that aspect of the situation receiving the observer's attention. In addition there is some work in the perception of language which indicates that the location of buzzes is often displaced to the beginning or end of certain kinds of psychological and linguistic units. Add to these the differences in reaction-time and difficulties in identifying brief pauses which occur at major syntactic breaks, and some variability in the exact temporal location of behavior codes is inevitable.

The measure of inter-observer agreement used was the Scott coefficient ($\pi = \frac{P_o - P_e}{1 - P_e}$, where P_o is the observed percent agreement and P_e is that expected by chance). This index takes account of the number of

categories in the system as well as the frequency with which each is used.

III. CLASSROOM SAMPLES

The sample of classroom observations was far from ideal. Teachers are very anxious when they suspect their teaching is being evaluated, even if informally and unofficially, and principals are at the moment loathe to even appear autocratic. Hence we had to be content with volunteers. In addition, the study used summer-session classes where there is much less pressure than in regular session, the pupils are in many cases volunteers, and there are many multi-grade classrooms. There were 6 schools, 20 teachers, 8 subject-matters, and 8 grade-groups; all told there were 33 sessions where the coding was independent. In some cases there were repeated measures on individual teachers. Most of the students came from homes of average socio-economic status; most were "Anglos" but there were some Mexican-Americans in many of the classes. The sessions varied in length from 7 to 34 minutes, a session being defined as a coherent curricular unit. Table I shows the distribution of sessions for the independent coding sessions.

IV. TRAINING OF CODERS

The two observers who participated in this study were both advanced graduate students in education. One had participated in the development of the instrument, and had considerable practice in coding cards bearing descriptions of single items of behavior, typescripts of classroom records, audio tapes, and a few video tapes, as well as a few hours of live classroom coding. The other observer had a crash course of about two weeks' duration under the guidance of the first, entirely in the laboratory setting. Their eleven non-independent coding sessions can be considered additional training.

V. RESULTS

The Scott coefficient (π) is shown in Table II for each independent session, in order of occurrence, together with the grand mean, and the means of successive thirds. The inter-observer agreement is about 75 percent. The upper curve in Fig. 1 shows that there is very little change over the course of the study, although the first 10 sessions are slightly less reliable.

The lower curve in Fig. 1 shows the total number of categories of behavior observed in each session; it is

relatively constant at a fairly high level. The mean number of categories coded is 22, with a range of 14 to 28 for single sessions. Table III shows the categories most frequently coded.

VI. DISCUSSION

The level of inter-observer agreement is moderately high, certainly high enough to be encouraging. Some of the sources of disagreement are known and are modifiable, either by more rigorous control of training or by modification of coding procedures. These are:

1. Omissions: These represented 18 percent of the coded behaviors. A large proportion of these referred to brief behaviors noted by one but not by the other observer: 1_1 (positive reinforcement) is often a perfunctory nod, "yes", "O.K.", or "right", immediately followed by a question or instruction which dominates the 5-second cell, so that the observer's attention is drawn to a more complex decision process; a_1 (constructive silence) is often brief, while awaiting a pupil's reponse, so one observer may judge it a normal pause, similar

to a syntactic break between speakers, whereas the other may think it somewhat longer, to give the pupil a brief time for thought.

2. Systematic Observer Bias: Among these biases were sensitivity to certain categories of behavior and insensitivity to others, different criteria for memory vs. thought processes or other distinctions, and differential knowledge of expected performance level of children of various grades (in judging whether a child's answer was likely to be memory or thought). We have no measures for these biases, but analysis of the nature of the confusions could suggest measures. Such biases were prominent in training, where every effort was made to reduce them to a minimum.

Other sources of disagreement are known but not easily dealt with. Among these are:

1. Audibility of teacher and pupil voices: This depends on ambient noise, classroom climate, acoustics, individual differences in voice and personality, and the observers' acuity. If things are bad enough, one can discard the

entire session, but ordinarily there are several spots where the message is unclear. Sometimes the observer with the better acuity hears it and the other does not; often it is missed by both.

2. Length of Observation: Short sessions are less reliable than long ones, but it is necessary to consider as a session only a coherent instructional unit. A lower limit should be set for an acceptable length. This should be established empirically, but would probably be between 20 and 30 minutes.

3. Use of a very small number of categories: This yields a high P_e, hence π will be low. This could occur in classes having special drill sessions, largely lecturing behavior on the part of the teacher, and rapid-fire question and answer session, etc.

4. Difficulty of synchronizing the timers of two observers: This might result in displacement of a code by as much as two cells, particularly if one observer habitually codes early in the period and the other later.

5. Rate of interaction in the classroom: Some teachers in certain situations move at a very rapid pace; question, call on student, answer, feedback, question, all within a 5-second interval. This can be hard to keep up with, especially if the decision about intellectual level of questions and answers is difficult. In rapid-fire situations, some events are inevitably omitted; also a change to a higher level of question after several lower level questions is likely to go unnoticed, unless an observer is particularly alert to that topic, and then one observer may be so alerted and the other not.

Finally there are the confusions which still occur because of the difficulty of defining the limits of the categories so that everyone interprets them in the same way. The categories in which the greatest amount of disagreement occurred were: $3_1, 3_2, 5_1, 5_4, 6_1, a_1$. a_1 is ordinarily very brief, as mentioned above. 3_1 and 3_2 could be defined more precisely, perhaps, but it is a matter of judgment whether the teacher is structuring the lesson or giving new material, a judgment peculiarly

difficult for a sudden visitor to make. It could be corrected by interviewing the teacher and observing for several days, but this is not a practical solution for an efficient evaluation instrument. It is difficult to see how 5_1 can be mistaken, but it is. Brief, immediate orders may be missed. It is also difficult to see how 6_1 can be mistaken, but there was, for these observers, some confusion with 6_2 and 7_1. It is also possible that some responses are so brief ("yes" or a nod), that they get lost like the other very brief ones do. All these categories occur in the most frequent class (see Table III), and the frequency of disagreement is not great in any session (see Table IV). As can be seen, for no category do any large number of sessions show disagreements; even for the categories where a relatively large proportion of disagreements occur in one session, other sessions typically show a small or insignificant proportion of disagreements.

Some categories never have any significant number of disagreements. These are: 2_1 (negative informational feedback), 6_5 ("I don't know"), 7_5 (irrelevant remarks), 8_1 (practice small unit), 8_2 (practice more complex unit), 9_1 (pupil-positive evaluation), 0_1 (pupil-negative

evaluation), a_3 (interruptions), a_4 (general noise and confusion), and a_5 (pupil's misbehavior). They are either quite clearly defined behavioral units such as 2_1 and a_3, or infrequent behaviors (6_5,7_5,8_1,8_2,9_1,0_1 and a_4). For categories with only a single instance of significant disagreement, the same holds true: 2_2, b_2, and b_5 are infrequent; 4_2 (thought questions) was the subject of much training and the decision criteria were made specific; b_1 (pupil's question or statement regarding procedure) was a clearly defined behavioral unit, not easily confused with others.

VII. CONCLUSIONS

The inter-observer agreement is sufficiently good to permit use of the instrument for evaluation purposes. The time necessary to train coders is not excessive; two weeks of half-time work seems adequate. It is probably not necessary to use graduate-level personnel, but, as an alternative some teaching experience would probably be necessary. Clerks would probably not be trainable in any reasonable length of time.

A modification of procedure should be introduced when the goal is evaluation of a school or a grade-level.

Instead of multiple codings in a single cell, a single code should be used, that for the behavior which is occurring at the time signal. If the session is sufficiently long (approximately 30 min.), the sampling of behavior at 5-second intervals will give a sufficiently accurate estimate of the important (i.e., frequently occurring) categories. In addition, a large source of disagreement (the very brief behaviors overshadowed by the more time-consuming ones) will be minimized, and the observers will not be subject to as much time stress as at present, itself a source of disagreement. There are special research and training problems for which the coding of behavior sequences is important, but for general evaluation purposes it is probably not necessary, and the suggested simplification will make both coding and analysis very much simpler.

REFERENCES

Amidon, E., Interaction analysis applied to teaching. National Association of Secondary School Principals' Bulletin, 1966, 50, (Dec.), 93-97.

Amidon, E.J. & Flanders, N.A., The effects of direct and indirect teacher influence on dependent prone students learning geometry. Journal of Educational Psychology, 1961, 52, 286-291.

Amidon, E.J. & Flanders, N.A., The role of the teacher in the classroom, (Rev. Ed.). Minneapolis: Association for Productive Teaching, 1967.

Anthony, B.C.M., The identification and measurement of classroom environmental process variables related to academic achievement. Ph.D. Thesis (Educ.), University of Chicago, 1967.

Bellack, A.A., Kliebard, H.M., Hyman, R.T., & Smith, F.L., Jr., The language of the classroom. New York: Teachers College Press, 1966.

Birkin, T.A., Toward a model of the instructional processes. Mimeo., University College of Townsville (Townsville, Queensland, Australia), 1968.

Bond, G.L. & Dykstra, R., The cooperative research program in first-grade reading instruction. Reading Research Quarterly, 1967, 2, No. 4 (pp. 142).

Boring, E.G., A history of experimental psychology, (2nd Ed.). New York: Appleton-Century-Crofts, 1950.

Flanders, N.A., Teachers influence, pupil attitudes and achievement. Mimeo., University of Michigan, Ann Arbor, 1962.

Gordon, C.W. & Adler, L.M., Dimensions of teacher leadership in classroom social systems: Effects on pupil productivity, morale, and compliance. Los Angeles: University of California, Cooperative Research Project, 1084 (DHEW), 1963.

Herman, W.L., An analysis of the activities and verbal behavior in selected fifth-grade social studies classes. Journal of Educational Research, 1967, 60, 339-345.

Perkins, M.V., Classroom behavior and underachievement. American Educational Research Journal, 1965, 2, 1-12.

Rosenshine, B., Teaching behaviors related to pupil achievement. Mimeo. 1969.

Schantz, B., An experimental study comparing the effects of recall by children in direct and indirect teaching methods as a tool of measurement. Doctoral Dissertation, Pennsylvania State University, 1963. (Dissertation Abstracts, 1963, 25, 1054).

Scott, W.A., Reliability of content analysis: A case of nominal scale coding. Public Opinion Quarterly, 1955, 321-325.

Simon, A. & Boyer, E.G., Mirrors for behavior: An anthology of classroom observation instruments, (6 vols.), Philadelphia: Research for Better Schools, 1967.

Taba, H., Levine, S., & Elzey, F., Thinking in elementary school children. Cooperative Research Project, 1574 (DHEW), San Francisco State College, 1964.

Weber, W.A., Teacher and pupil creativity. Doctoral Dissertation, Temple University, 1967. (Dissertation Abstracts, 1967, 29, 159A).

FIGURE 1 - SACC

Inter-Observer Agreement

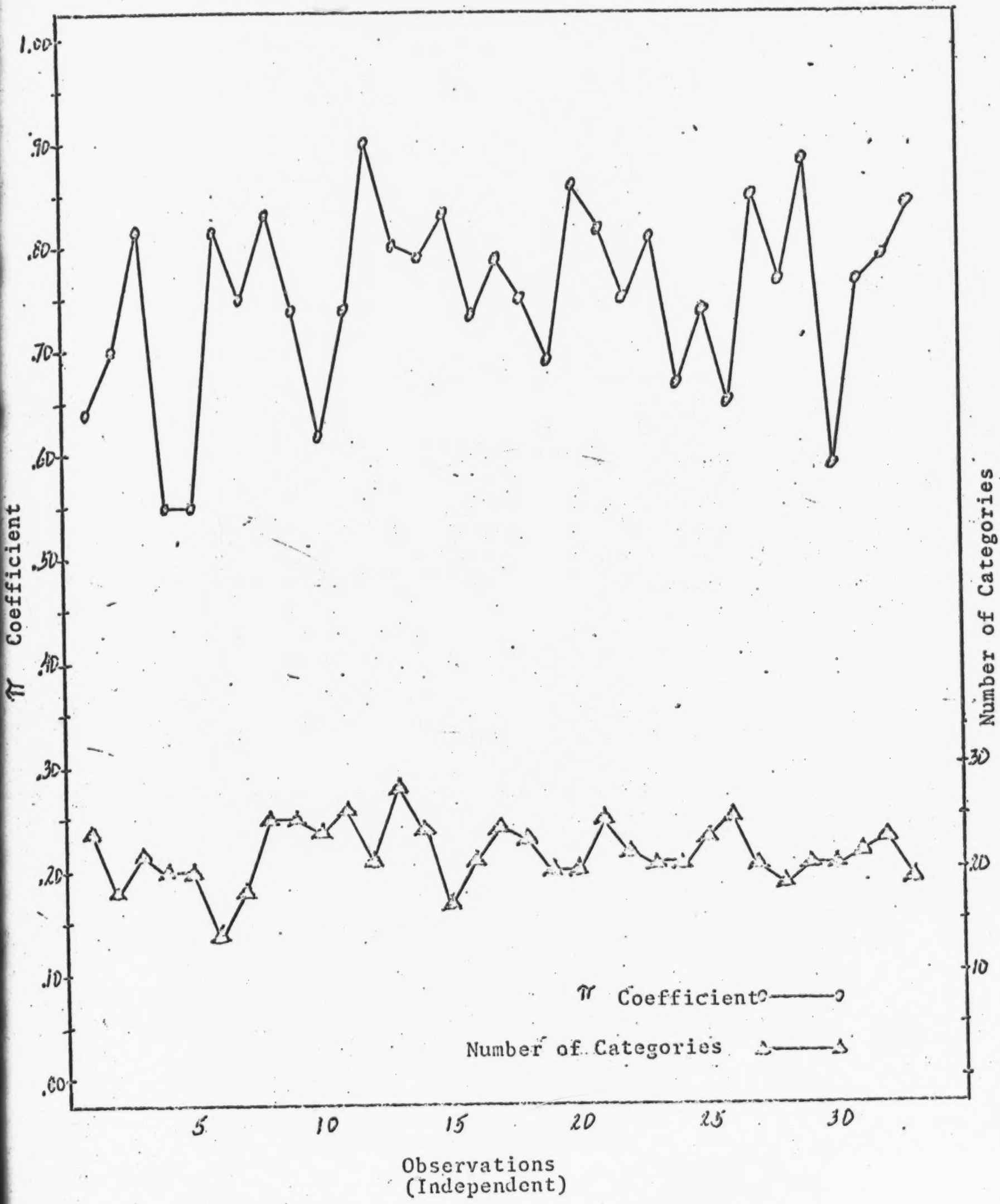


Table 1: Distribution of Coding Sessions

Subject	N	GRADE 1-2	2	1-2-3	2-3	3-4	4-5	5-6	6	7	8	7-8	2-5	4-6
Social Studies	4			1				1		2				
Reading	4	3				1								
Math	9	1			1	1	5					1		
Language Arts	8	5							1				1	1
Science	5		1			2		1				1		
Fire Safety	1					1								
Art	1											1		
Foreign Language	1			1										

Table II

Inter-observer Agreement for SACC, Form V*

servation	π	Observation	π	Observation	π
1	.64	12	.90	23	.81
2	.70	13	.80	24	.67
3	.82	14	.79	25	.74
4	.55	15	.83	26	.65
5	.55	16	.73	27	.85
6	.82	17	.79	28	.77
7	.75	18	.75	29	.88
8	.83	19	.69	30	.59
9	.74	20	.86	31	.77
10	.62	21	.82	32	.79
11	.74	22	.75	33	.84

Total:

$\bar{\pi} = .75$

$R_{\pi} = .55\text{-}.90$

First 10 Observations:

$\bar{\pi} = .70$

Second 10 Observations:

$\bar{\pi} = .79$

Last 13 Observations:

$\bar{\pi}$ .76

TABLE III

Observed Frequency of Categories*

Most Coded	Moderately Coded	Least Coded	Rarely Coded
1_1	1_2	2_2	6_5
2_3	2_1	7_2	a_5
3_1	4_2	7_5	
3_2	6_2	8_1	
4_1	7_1	8_2	
5_1	a_3	9_1	
5_2	b_1	0_1	
5_3		a_4	
5_4		b_2	
6_1		b_5	
a_1			
a_2			

Most: $\chi \geq 15\%$

Moderate: $5\% \leq \chi < 15\%$

Least: $1\% \leq \chi < 5\%$

Rare: $\chi < 1\%$

Average % of codings per observation for two observers

*Independent sessions only

TABLE IV DAILY % DISAGREEMENT BETWEEN RATERS BY CATEGORY (72.3%) FOR 44 SESSIONS (INDEPENDENT AND NON-INDEPENDENT)

Category	6/27	6/30	7/1	7/2	7/3	7/7	7/8	7/11	7/15	7/16	7/17	7/18	7/22	7/23	7/25	7/28	N SESSIONS WITH DISAGREEMENTS
1	5.6	3.6	3.2	3.1	3.1	2.6	2.4	4.0	4.4						3.7		8
2		3.2	3.2														2
1																	0
2			2.4														1
3			4.5				2.5			3.1							3
1	5.8	3.9	5.5	7.2	8.0	3.3	2.9	2.8			2.6		6.1	2.9	2.8		12
2	11.3	5.8	4.9	10.5	4.1	2.9	3.6	5.3	2.6		2.4	2.4		9.2			12
1	3.2			3.7	2.9	4.2	13.8	2.6		2.6	3.9		4.6		2.4		10
2													6.5				1
1	10.0	2.9	2.8		4.2	2.7	3.1	2.5	7.2	4.8						5.3	10
2	3.3	2.4			2.4	3.6							2.5	5.7			6
3		2.7				3.9		2.8		2.6	2.6	2.8					6
4				2.5	11.1	2.6											3
1	9.9	2.4	2.4	2.4	5.5	3.9				6.2						2.4	8
2							3.1			3.3					6.6		3
5																	0
1		2.8	5.6			3.3		3.4	2.5								5
2			3.2									3.7			2.7		3
5																	0
1																	0
2																	0
1																	0
1																	0
1	3.4	14.4		2.8	8.6	3.7	3.0	2.8	2.5	2.4	4.5		7.4		9.8		13
2			2.6													2.4	2
3															3.8		0
4																	0
1		2.7															1
2																2.6	1
5	6.1																1
	6	4	3	7	5	4	2	3	5	2	1	3	4	7	5	4	1	4	2	3	5	1	1	1	2	3	4	1	3	2	2	1	2	1	1	3	1	2	1	

APPENDIX*

* Prepared by Margaret Hubbard Jones, with the assistance of E.M. Swengel, M.K. Osman, and K. Olivier, July, 1969.

C.S.E. System for the Analysis of Classroom Communication (SACC)-Form V (Short Form)

A - Teacher's Behavior	1	2	3	4	5
1 - Positive Affect	Positive reinforcement, encouragement in cognitive task.	Positive reinforcement: T's admission of error, acknowledgment of P's feelings; social chit-chat.			
2 - Negative Affect	Negative informational feedback; correctional response.	T's justification of judgment or decision.	Personal criticism; threat or punishment.		
3 - Teacher's Statements	Structures material; summarizes P's remarks.	Factual information, Yes, informational response; demonstrations, reads aloud; explanations, shows relationships, implications.			
4 - Questions: answers to be based on:	Memory: rote response; naming; preference (w/out criteria); meaning of term, definition; facts.	Thinking; evaluation, judgment, opinion with criteria; combination of several items of information.			
5 - Control of Behavior: classroom management	Assignment; check on assignment, directions for doing assignment. Immediate orders for work ("Open books;" "Put away...").	Calls on P to recite, do cognitive task.	Directs non-cognitive operations: "Shut the door;" run errand; get lunch boxes; line up.	T calls P's name.	
B.- Pupils' Behavior					
6 - Pupil Responses	Memory: rote response; naming; preference or opinion w/out criteria; meaning of term, definition; recital of text.	Thinking: convergent, divergent; evaluation, judgment, opinion with criteria, response showing processing of information.			"I don't know et al.